Numerical Procedures for Estimating the Parameters in a Multivariate Homogeneous Correlation Model with Unequal Variances

Juan C. Meza [*] and Ingram Olkin [†]

Sandia National Laboratories

Stanford University

Abstract

Closed form expressions for the maximum likelihood estimates (MLE) of the parameters in a multivariate normal distribution in which the variances are homogeneous and the correlations are equal is well-known. However, when the correlations are equal but the variances are not homogeneous, no closed-form expressions are available. We provide two iterative procedures that converge rapidly. One procedure uses an extension of a well-known scaling method, which is itself of intrinsic interest.

Keywords: maximum likelihood estimation, multivariate analysis, intraclass model

AMS Classifications:62H12, 62U05

[*] Supported in part by the Applied Mathematical Science Program, U.S. Department of Energy, Office of Energy Research and was performed at Sandia National Laboratories operated for the U.S. Department of Energy under contract No. DE-AC04-94AL85000.

[†] Supported in part by the National Science Foundation under grant 123456789.

Numerical Procedures for Estimating the Parameters in a Multivariate Homogeneous Correlation Model with Unequal Variances

Juan C. Meza [1] and Ingram Olkin [2]

Sandia National Laboratories Stanford University

1. Introduction

With the underlying assumption of a multivariate normal distribution, the intraclass covariance matrix $\Sigma = (\sigma_{ij})$, with homogeneous variances $\sigma_{11}^2 = \ldots = \sigma_{pp}^2 \equiv \sigma^2$ and homogeneous correlations $\rho_{ij} = \rho$ has been used to model phenomena that are permutation invariant, as for example, in the case of p equivalent psychological tests. It was first studied by Wilks (1946) and extended by Votaw (1948) to a more general model in which there is permutation invariance within blocks. Other multivariate extensions of the intraclass model have been considered by Olkin (1974), and estimates of the parameters by methods other than maximum likelihood estimates (MLE) have been obtained (see for example, Rothblum and Schneider (1989)).

In some applications the variances may differ, yet the correlations remain homogeneous. This modification has a profound effect on the maximization procedures. Whereas for a normal sample the MLE of σ^2 and ρ in the intraclass model are expressible in closed form, this is no longer the case when the variances are unequal. This paper provides an examination of two new numerical procedures for obtaining maximum likelihood estimates in the unequal variance case. One of the procedures requires an extension of a scaling theorem that has some intrinsic interest.

To set forth our notation, let the n rows of

$$X = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ldots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix},$$

denote independent observations from a p-dimensional random variable having a normal distribution with mean vector zero and covariance matrix Σ. Then $S = (XX')$ is the cross-product matrix having a Wishart distribution $\mathcal{W}(\Sigma; p, n)$ with density function

$$f(S) = c_0|\Sigma|^{-n/2}|S|^{(n-p-1)/2}\exp\left[-\frac{1}{2}\text{ tr } S\Sigma^{-1}\right], \quad S > 0, \Sigma > 0, \tag{1.1}$$

where c_0 is a normalizing constant, and $A > 0$ means that the matrix A is positive definite. When Σ is the intraclass covariance matrix

$$\Sigma_I = \sigma^2 P \equiv \sigma^2 \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \ldots & \rho & 1 \end{bmatrix}, \quad -1/(p-1) \leq \rho \leq 1, \sigma > 0, \tag{1.2}$$

the MLE of σ^2 and ρ are well-known (Wilks, 1946):

$$\hat{\sigma}^2 = \frac{\text{tr } V}{p}, \qquad \hat{\rho} = \frac{eVe' - \text{ tr } V}{(p-1)\text{ tr } V}, \tag{1.3}$$

where $e = (1, \ldots, 1), V = S/n$.

When the variances are unequal, we denote the covariance matrix by

$$\Sigma_{IV} = D_\sigma P D_\sigma, \quad D_\sigma = diag(\sigma_1, \ldots, \sigma_p), \tag{1.4}$$

where the designation IV refers to an intraclass correlation matrix coupled with unequal variances. Because Σ^{-1} occurs in (1.1), we let $D_\sigma^{-1} \equiv D_\tau = diag(\tau_1, \ldots, \tau_p)$. Denote the Hadamard or Schur product by $A \circ B = (a_{ij}b_{ij})$.

2. An iterative ML Method

To motivate the iterative procedure, rewrite (1.1) when $\Sigma = \Sigma_{IV}$ given by (1.4):

$$f(S) = c(S)(\prod_{i=1}^{p} \tau_i^n)|P|^{-n/2}\exp -\frac{1}{2}\left[\text{ tr } SD_\tau P^{-1} D_\tau\right], \tag{2.1}$$

where $c(S) = c_0|S|^{(n-p-1)/2}$. Let $Q = (V \circ P^{-1})$. Then (2.1) can be rewritten as

$$[f(S)]^{\frac{1}{n}} = [c(S)]^{\frac{1}{n}}(\prod_{i=1}^{p} \tau_i)|P|^{-1/2}\exp -\frac{1}{2}(\underset{\sim}{\tau} Q \underset{\sim}{\tau}'), \tag{2.2}$$

where $\underset{\sim}{\tau} = (\tau_1, \ldots, \tau_p)$.

To obtain the MLE of Σ_{IV}, we need to resolve the maximization problem

$$\max_{\rho,\, \underset{\sim}{\tau}} \; (\prod_{i=1}^{p} \tau_i)|P|^{-1/2} \exp -\frac{1}{2}(\underset{\sim}{\tau} Q \underset{\sim}{\tau}'), \tag{2.3}$$

noting that Q is a function of ρ. Our procedure is to first maximize with respect to $\underset{\sim}{\tau}$ using a fixed ρ, then maximize with respect to ρ using a fixed $\underset{\sim}{\tau}$, and iterate. Maximizing (2.3) with respect to $\underset{\sim}{\tau}$ ($\tau_i > 0, i = 1, \ldots, p$) is equivalent to maximizing

$$\sum_{i=1}^{p} \log \tau_i - \frac{1}{2} \underset{\sim}{\tau} Q \underset{\sim}{\tau}', \quad \underset{\sim}{\tau} > 0, \tag{2.4}$$

which is a strictly concave function of $\underset{\sim}{\tau}$. The first derivative equation is

$$\left(\frac{1}{\tau_1}, \ldots, \frac{1}{\tau_p} \right) = (\tau_1, \ldots, \tau_p)\, Q, \tag{2.5}$$

which can be written as $eD_\tau^{-1} = eD_\tau Q$, or as $e = eD_\tau Q D_\tau$.

This problem can also be posed as maximizing a convex function subject to a constraint:

$$\max_{x} \; x \; Q \; x' \quad \text{subject to } \prod x_i = 1, x_i > 0,$$

or more conveniently, subject to $\sum \log x_i = 0$, $x_i > 0$.

The solution to (2.5) is the solution to a particular scaling result of Sinkhorn(1964), namely, given a positive definite matrix Q, there exists a diagonal matrix D_τ with positive diagonal elements such that $D_\tau Q D_\tau$ is doubly stochastic. This problem has a considerable history, see for example, Marshall and Olkin (1968). In particular, there are numerical procedures for solving (2.5) or for finding the maximum in (2.4).

Thus, given ρ, which means that Q is given, we can determine the maximizing vector $\underset{\sim}{\tau}$. Now consider the density (2.1) in terms of ρ. Note that

$$P^{-1} = aI + bee', \quad a = 1/(1-\rho), \quad b = -\rho/[(1-\rho)(1+(p-1)\rho)]. \tag{2.6}$$

Furthermore, $a + pb = [1 + (p-1)\rho]^{-1}$, and $|P|^{-1} = a^{p-1}(a + pb)$. In (2.1) let $U = D_\tau V D_\tau$, so that the maximization with respect to ρ becomes

$$\begin{aligned} &\max_{\rho} \quad |P|^{-1/2} \exp -\frac{1}{2}(\text{ tr } UP^{-1}) = \\ &\qquad \max_{\rho} \quad [a^{p-1}(a+pb)]^{1/2} \exp -\frac{1}{2}(u_1(a+b) + u_2 b), \end{aligned} \tag{2.7}$$

where $u_1 = \sum u_{ii}, u_2 = \sum_{i \neq j} u_{ij}$. Taking logarithms in (2.7), the first derivative equation becomes

$$\frac{p-1}{1-\rho} - \frac{p-1}{1+(p-1)\rho} - u_1(\frac{\partial a}{\partial \rho} + \frac{\partial b}{\partial \rho}) - u_2 \frac{\partial b}{\partial \rho} = 0, \tag{2.8}$$

where

$$\frac{\partial a}{\partial \rho} = \frac{1}{(1-\rho)^2}, \qquad \frac{\partial b}{\partial \rho} = \frac{-(1+(p-1)\rho^2)}{(1-\rho)^2[1+(p-1)\rho]^2}.$$

Equation (2.8) is a cubic equation, which can readily be solved numerically to obtain the maximizer.

Thus, we have that given ρ we can solve for $\underset{\sim}{\tau}$ and vice versa. This suggests the following algorithm:

Algorithm 1.

Step 0. Choose an admissible ρ^0. For example, choose ρ^0 to be the solution under the intraclass covariance model.

Step 1. Use (2.5) with ρ^0 to obtain τ^1. Note that $Q = (q_{ij})$, where $q_{ii} = v_{ii}(a+b), q_{ij} = v_{ij}, i \neq j$, and a and b are defined in terms of ρ^0 in (2.6).

Step 2. Using the updated τ^1 and (2.8) find ρ^1. Note that u_1 and u_2 are defined in terms of τ^0.

Step 3. Iterate until the process converges.

In this process, each iteration consists of the solution of a Sinkhorn scaling problem (in Step 1) and the computation of the roots of a cubic equation (in Step 2).

3. An Alternative Iterative ML Method

By a reparameterization we can take advantage of the fact that the MLE for the intraclass covariance model can be obtained explicitly. We first review the ingredients needed later. Rewrite the model (1.4) as

$$\Sigma_{IV} = D_\sigma P D_\sigma = D_\eta \Sigma_I D_\eta, \tag{3.1}$$

where $D_\eta \equiv diag(1, \eta_2, \cdots, \eta_p) = diag(1, \frac{\sigma_2}{\sigma_1}, \cdots, \frac{\sigma_p}{\sigma_1})$ and $\Sigma_I = \sigma_1^2 P$. For simplicity we write σ^2 for σ_1^2.

Recall that the characteristic roots of Σ_I are $\alpha \equiv \sigma^2[1 + (p-1)\rho]$ and $\beta \equiv \sigma^2(1-\rho)$ of multiplicity $p-1$. Consequently,

$$|\Sigma_{IV}| = (\prod_2^p \eta_i^2)\ \alpha\beta^{p-1}. \tag{3.2}$$

Then (2.2) becomes

$$[f(S)]^{1/n} = [c(S)]^{1/n} (\prod_2^p \eta_i)\ (\alpha\beta^{p-1})^{-1/2} \exp -\frac{1}{2}(\text{ tr } V D_\eta^{-1} \Sigma_I^{-1} D_\eta^{-1}). \tag{3.3}$$

Further, any orthogonal matrix G with first row $e/\sqrt{p}$ diagonalizes Σ_I, that is, for any such orthogonal G, $G\Sigma_I G' = D = diag(\alpha, \beta, \ldots, \beta)$. Consequently, with $W = D_\eta^{-1} V D_\eta^{-1}$,

$$\begin{aligned} \text{tr } W\Sigma_I^{-1} &= \text{tr } (GWG')(G\Sigma_I^{-1}G') \\ &= \text{tr } \widetilde{W} D^{-1} = \frac{\widetilde{w}_{11}}{\alpha} + \frac{\sum_2^p \widetilde{w}_{ii}}{\beta}, \end{aligned} \tag{3.4}$$

where $\widetilde{W} = GWG'$. Thus, under the intraclass correlation model, a direct computation using the result (1.3) yields the MLE:

$$\hat{\alpha} = \hat{\sigma}^2[1 + (p-1)\hat{\rho}] = \frac{e\widetilde{W}e'}{p}, \qquad \hat{\beta} = \hat{\sigma}^2(1-\hat{\rho}) = \frac{p\ tr\widetilde{W} - e\widetilde{W}e'}{p(p-1)}, \tag{3.5}$$

from which

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\hat{\alpha} + (p-1)\hat{\beta}}{p} = \frac{tr\widetilde{W}}{p}, \\ \hat{\rho} &= \frac{\hat{\alpha} - \hat{\beta}}{\hat{\alpha} + (p-1)\hat{\beta}} = \frac{e\widetilde{W}e' - tr\widetilde{W}}{(p-1)tr\widetilde{W}}. \end{aligned} \tag{3.6}$$

We can now make use of (3.5) in the more general model. From (1.1) using (3.1) and (3.2):

$$\begin{aligned} [f(S)]^{1/n} &= c(S)^{1/n} \left[(\prod_2^p \eta_i^2)\ \alpha\beta^{p-1} \right]^{-1/2} \exp\left[-\frac{1}{2} trV D_\eta^{-1} G' (G\Sigma_I^{-1} G') G D_\eta^{-1} \right], \\ &= c(S)^{1/n} \left[(\prod_2^p \eta_i^2)\ \alpha\beta^{p-1} \right]^{-1/2} \exp\left[-\frac{1}{2} tr(G D_\eta^{-1} V D_\eta^{-1} G') D^{-1} \right], \end{aligned} \tag{3.7}$$

where $D = diag(\alpha, \beta, \ldots, \beta)$. For fixed η, let $\widetilde{W} = GD_\eta^{-1}VD_\eta^{-1}G'$ in which case (3.7) becomes

$$[f(S)]^{1/n} = c(S)^{1/n} \left[(\prod_2^p \eta_i^2)\ \alpha\beta^{p-1} \right]^{-1/2} \exp -\frac{1}{2}(\text{ tr } \widetilde{W}D^{-1}),$$

so that $\hat{\alpha}$ and $\hat{\beta}$ are given by (3.5). Recall that

$$\text{tr } D_\eta^{-1}VD_\eta^{-1}H = (1, \underset{\sim}{\nu})\ T\ (1, \underset{\sim}{\nu})', \quad \underset{\sim}{\nu} = (\nu_2, \ldots, \nu_p),$$

where $\nu_i = 1/\eta_i$, $i = 2, \ldots, p$, and $T = V \circ H$. Consequently, for fixed α and β, and $H = G'D^{-1}G$ we need to maximize

$$\sum_2^p \log \nu_i - \frac{1}{2}(1, \underset{\sim}{\nu})T(1, \underset{\sim}{\nu})'. \tag{3.8}$$

If we partition the matrix T such that

$$T = \begin{bmatrix} t_{11} & \underset{\sim}{t} \\ \underset{\sim}{t}' & \widetilde{T} \end{bmatrix},$$

where $\underset{\sim}{t} = (t_{12}, \ldots, t_{1p})$, then (3.8) becomes

$$\sum_2^p \log \nu_i - \frac{1}{2}(t_{11} + 2\underset{\sim}{\nu}\underset{\sim}{t}' + \underset{\sim}{\nu}\widetilde{T}\underset{\sim}{\nu}'). \tag{3.9}$$

The first derivative equation of (3.9) is

$$\left(\frac{1}{\nu_2}, \ldots, \frac{1}{\nu_p} \right) = \underset{\sim}{t} + \underset{\sim}{\nu}\widetilde{T}. \tag{3.10}$$

When $\underset{\sim}{t} = 0$, (3.10) reduces to (2.5), which is the Sinkhorn problem. The essence of the maximization has not changed in that $(x + a)\ Q\ (x + a)'$ is a convex function of x. Thus we need to determine

$$\max_x\ (x + a)\ Q\ (x + a)',\ a \text{ an arbitrary vector},$$

subject to $\prod x_i = 1, x_i > 0$, or to $\sum \log x_i = 0, x_i > 0$ which is a convex constraint. This leads to an alternative algorithm:

Algorithm 2.

Step 0. Choose admissible initial values α^0 and β^0, for example, $\alpha^0 = \beta^0 = 1$.

Step 1. Solve for $\underset{\sim}{\nu}^0$ by either (3.9) or (3.10).

Step 2. Using $\underset{\sim}{\nu}^0$ and (3.5) compute a new α^1 and β^1.

Step 3. Iterate until the process converges.

4. Numerical Results

To test out the new algorithms we ran several numerical experiments on some model problems. All computer runs were made on an SGI Indigo workstation in double precision arithmetic. As a benchmark, we used the nonlinear optimization package NPSOL developed by Gill, Murray, Saunders, and Wright (1986).

We were particularly interested in the effects of the dimension of the problem and the value of the correlation coefficient. We constructed a model problem by computing a cross product matrix from a set of random variables generated from a normal distribution with mean zero and a $p \times p$ covariance matrix given by

$$\Sigma_{IV} = D_\sigma P D_\sigma, \quad D_\sigma = diag(\sigma_1, \ldots, \sigma_p), \quad P = (1-\rho)I + \rho e e',$$

where $D_\sigma = diag(\sqrt{0.7}, \sqrt{1.5}, \sqrt{2.0}, \sqrt{2.3}, \sqrt{3.0}, \sqrt{0.7}, \ldots, \sqrt{3.0})$, and $\rho = 0.6, 0.9$. For each value of ρ, we ran four test problems with $p = 5, 20, 50, 100$. The number of observations, n, was set equal to $10p$.

For the method NPSOL, the convergence tolerance was set to 10^{-12}, which according to the user's guide is a rough approximation to the number of correct figures desired in the objective function at the solution. For the new algorithms, the methods were terminated if any of the following three conditions were met:

$$(4.1) \qquad ||g_{k+1}|| \leq 10^{-6}(1 + |f_{k+1}|),$$

$$(4.2) \qquad ||x_{k+1} - x_k|| \leq 10^{-9}||x_{k+1}||,$$

$$(4.3) \qquad ||f_{k+1} - f_k|| \leq 10^{-9}(1 + |f_{k+1}|),$$

where f_k and g_k are the function and gradient values respectively at the k-th iteration.

Condition (4.1) is the most satisfactory termination criteria in that the gradient will be approximately zero. The other two conditions are necessary to keep the algorithms from taking too many iterations in regions where insufficient progress is being made. The tolerances were chosen to correspond approximately with those used by NPSOL.

The results from using NPSOL on the test problems are displayed in Table 4.1. The table contains the value of the likelihood function and the norm of the gradient at the solution in columns 3-4. In addition, the number of iterations, the total number of function evaluations and the total cpu time taken are shown as a way of comparing the relative efficiency of the various algorithms. Tables 4.2-4.3 contain the numerical results from algorithms 1 and 2 respectively.

Table 4.1: NPSOL

p	ρ	$f(x^*)$	$\|\|g(x^*)\|\|$	Iter(Feval)	Cpu time
5	0.6	2.760	$7.28\ 10^{-7}$	11(19)	0.05
20	0.6	7.904	$3.37\ 10^{-6}$	23(49)	0.19
50	0.6	17.439	$4.56\ 10^{-6}$	25(53)	0.82
100	0.6	33.322	$4.49\ 10^{-6}$	28(58)	3.34
5	0.9	0.139	$1.99\ 10^{-7}$	17(36)	0.05
20	0.9	-5.084	$9.04\ 10^{-7}$	32(74)	0.26
50	0.9	-16.322	$1.01\ 10^{-5}$	38(84)	1.31
100	0.9	-35.111	$3.27\ 10^{-5}$	45(103)	5.84

Our first observation is that all of the methods return the same likelihood value in every test case. We also note that in most cases, both of the new algorithms yield a solution with fewer function evaluations (as well as faster execution times) than NPSOL. Algorithm 1 also gives solutions that have roughly the same amount of accuracy as the solutions returned

Table 4.2: Algorithm 1

p	ρ	f	$\|\|g\|\|$	Iter(Feval)	Cpu time
5	0.6	2.760	$1.96\ 10^{-8}$	8(24)	0.11
20	0.6	7.904	$2.91\ 10^{-5}$	8(31)	0.23
50	0.6	17.439	$5.34\ 10^{-5}$	8(31)	0.76
100	0.6	33.322	$5.68\ 10^{-5}$	8(29)	2.29
5	0.9	0.139	$6.29\ 10^{-8}$	11(43)	0.15
20	0.9	-5.084	$1.89\ 10^{-5}$	11(44)	0.29
50	0.9	-16.322	$1.65\ 10^{-5}$	10(39)	0.97
100	0.9	-35.111	$2.62\ 10^{-5}$	10(39)	2.48

Table 4.3: Algorithm 2

p	ρ	f	$\|\|g\|\|$	Iter(Feval)	Cpu time
5	0.6	2.760	$8.04\ 10^{-6}$	9(26)	0.10
20	0.6	7.904	$9.95\ 10^{-4}$	8(32)	0.24
50	0.6	17.439	$3.86\ 10^{-3}$	8(32)	0.95
100	0.6	33.322	$2.12\ 10^{-3}$	8(29)	3.09
5	0.9	0.139	$5.60\ 10^{-4}$	7(30)	0.11
20	0.9	-5.084	$1.09\ 10^{-3}$	9(27)	0.22
50	0.9	-16.322	$1.04\ 10^{-3}$	11(45)	1.15
100	0.9	-35.111	$1.17\ 10^{-3}$	11(41)	4.18

from NPSOL as measured by the norm of the gradients. The second algorithm is as efficient as the first algorithm but always returned a solution with a larger gradient.

An interesting point is that the dimension of the problem did not affect the convergence of either of the new algorithms. In all cases, the number of iterations was between 8-11 iterations. In contrast, the number of iterations required by NPSOL grows with the dimension of the problem. The number of iterations required by NPSOL also increased as the value of the correlation coefficient was increased. In the largest dimensional problem the number of function evaluations required almost doubled. The two new algorithms were not substantially affected by the larger correlation coefficient.

A final point concerns the choice of the initial guess for the two new algorithms. Although there are clear choices in each case (for example the solution under the intraclass covariance model for Algorithm 1), we also tested the algorithms over a wide range of initial guesses. In particular, for Algorithm 1 we also conducted tests using an initial guess for ρ close to both $-1/(p-1)$ and 1, which constitutes the extreme values for this parameter. For Algorithm 2, the extreme case consists of choosing $\alpha = \beta = 0$. In all the numerical tests using these initial guesses both algorithms converged. Interestingly enough, in some cases the algorithms converged in a fewer number of iterations than the cases displayed in Tables 4.1-4.2, although on the average the use of extreme values for the initial guess caused both algorithms to take a larger number of iterations.

5. Conclusions

We have presented two new algorithms for the computation of maximum likelihood estimators in the case of multivariate normal distributions where the correlations are equal but the variances are not homogeneous. The first algorithm involves the solution of a Sinkhorn scaling problem in conjunction with the solution of a cubic equation. The second algorithm only requires the solution of an extended Sinkhorn problem which is of intrinsic interest. Both new algorithms compare favorably with standard nonlinear programming techniques. In addition, the new algorithms have the advantage of being fairly insensitive to the dimension of the problem and to the correlation coefficient. These two characteristics make the new algorithms more attractive for large dimensional problems.

References

[1] Gill, P., Murray, W., Saunders, M., and Wright, M.(1986). User's guide for NPSOL (Version 4.0): A FORTRAN package for nonlinear programming, *Stanford University Technical Report SOL 86-2.*

[2] Marshall, A.W. and Olkin, I. (1968). Scaling of matrices to achieve specified row and column sums, *Numerische Mathematik*, 12, 83-90.

[3] Olkin, I. (1974). Inference for a normal population when the parameters exhibit some structure, pp. 759-773 in *Reliability and Biometry: Statistical Analysis of Lifelength*, (ed. by F. Proschan and R.J. Sertling), SIAM: Philadelphia.

[4] Rothblum, U.G. and Schneider, H. (1964). Scalings of matrices which have prespecified row sums and column sums via optimization, *Lin. Alg. Appl.*, 114/115, 737-764.

[5] Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices, *Ann. Math. Statist.*, 35, 876-879.

[6] Votaw, D.F. Jr. (1948). Testing compound symmetry in a normal multivariate distribution, *Ann. Math. Statist.*, 19, 447-473.

[7] Wilks, S.S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution, *Ann. Math. Statist.*, 17, 257-281.